DOCUMENT RESUME

ED 097 336 TM 003 861

AUTHOR	Flaugher, Ronald L.
TITLE	The New Definitions of Test Fairness in Selection: Developments and Implications. Research Memorandum No. 73-17.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RM-73-17
PUB DATE	Sep 73
NOTE	15p.
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	Models; *Selection; *Test Bias; *Testing Problems

ABSTRACT

Complexities of test fairness are described in nontechnical language, and their implications for the selection procedures practiced in our society are discussed. Four clearly distinguishable models of fair selection are presented: the Cleary, or traditional, model; the Cole model; the Thorndike model; and the Darlington model. A distinction is made between the use of tests in a manner which is "fair", and the concept of "test bias", which most frequently refers to the content of the items of the test, regardless of any particular use to which the test is being put. It is possible to conceive of a biased test being used in a fair manner, and also possible to imagine an unbiased test being used unfairly. This discussion concerns the use, rather than the content, of tests. (RC)

RM-73-17

RESEARCH MEMORANDUM

THE NEW DEFINITIONS OF TEST FAIRNESS IN SELECTION: DEVELOPMENTS AND IMPLICATIONS

Ronald L. Flaugher

This Memorandum is for interoffice use. It is not to be cited as a published report without the specific permission of the author.

Educational Testing Service
Princeton, New Jersey
September 1973

The New Definitions of Test Fairness in Selection: Developments and Implications[1]

Ronald L. Flaugher

Until recent years the topic of test fairness in selection was not a controversial one for those familiar with the psychometric theory underlying it. There was criticism of the use of tests as selection devices, but almost exclusively from a nontechnical standpoint, offering no effective challenge to the supposedly immutable scientific understanding of the problem of fairness.

With the publication of the Journal of Educational Measurement in the summer of 1971, the complacency of the theorists was shaken. In two articles, one by Thorndike and one by Darlington, the traditional model of test fairness was seen very clearly to be a great deal less immutable than had been supposed, and new models of fairness were presented that not only competed with but were incompatible with the traditional model and with each other, except in the impossible circumstance of a perfect correlation between a test score and the criterion performance.

The theory of test fairness was shown to be a complex issue, a discovery that had immediate real-life implications in a time of increased attention to the fair treatment of minority groups and concern with their selection for employment and education. In the remainder of this paper, an attempt will be made to describe the complexities in nontechnical language and to discuss their implications for the selection procedures practiced in our society.

---

[1]The author is grateful to Robert L. Linn, Joel T. Campbell, Charles W. Daves and Lewis W. Pike for helpful criticisms of an earlier version of this paper.

It is essential to this entire discussion to be aware of the distinction between the use of tests in a manner which is "fair," and the concept of "test bias," which most frequently refers to the content of the items of the test, regardless of any particular use to which the test is being put. It is possible to conceive of a biased test being used in a manner which is fair (by awarding bonus points, for example), and also possible to imagine an unbiased test being used unfairly (by using a test that is unrelated to the task being predicted). The present discussion concerns the use, rather than the content, of tests.

For the present purposes, there are just four clearly distinguishable models of fair selection, which can be named the Cleary, or traditional, model, the Thorndike model, the Cole model, and the Darlington model. The Cleary model, so named because of its use by Cleary (1968) in one of the first formal investigations of the fairness of tests in selection, is that model which was most widely accepted until the recent developments. Very simply, it states that a test is fair for both of two subgroups of a population (for example, an ethnic minority and majority group) if the prediction equation that is used neither systematically overpredicts nor underpredicts the level of performance for either group. This seems intuitively fair, in that underprediction is the frequent charge made against tests when used in the selection of minority group members: the accusation is that the test gives an artificially depressed estimation of the minority person's capacity to perform on the job, and if permitted to attempt the job, many would succeed who were predicted to fail. A fair test, on the other hand, is viewed as one which does not give this inaccurate picture of minority group members.

With the appearance of the Thorndike and Darlington articles, the use of this definition was seen to have a flaw that resulted in a situation which most people would agree was obviously unfair when it was applied to certain frequently occurring real-life situations, even though remaining fair by the previous standards. The reasons for this paradox are rather complex, but can be stated briefly in this way: It is often the case that the average test score for a minority group is somewhat lower than that for the majority group; further, on a criterion such as rated performance on the job, there is frequently a difference in the same direction, but the difference is somewhat less.[2] Thorndike pointed out in his article that under this very common circumstance, and even when the procedure meets the Cleary definition of fairness, a distinctly smaller percentage of the minority applicants will be selected than would have succeeded on the job. Another way of saying this is that if all minority applicants had been hired, then (for example) a third of them would have succeeded on the job; if all the majority applicants had been hired, then perhaps half would have succeeded on the job. However, under such circumstances it is entirely possible that one-half of the majority, but none of the minority group, would actually be selected by the traditional procedures. Thorndike very reasonably suggested that the traditional or Cleary definition of fairness is therefore unacceptable.

The alternative model that Thorndike suggested is directed specifically to that inequity of selection-versus-success for two subgroups of a population. He suggested that the base-rate of success for the various groups be determined

---

[2]This reduced difference on the criterion may be due to less reliability in its measurement: supervisor's ratings or academic grades, which are typically less reliable than objective standardized tests, are frequently used as criterion measures.

empirically, then that percentage of the applicant group be selected, whether or not the same cut-off score on the test is used for each of the groups. For example, if one-third of the minority group is found to be likely to succeed on the job, then the cut-off score for selection should be adjusted to hire one-third of that applicant group (or whatever proportion, relative to the other group, is permitted by the number of openings). The traditional method may be fair for the individual, claims Thorndike, but in order for the method to be fair for the group, his method must be used.

The impact of Thorndike's article on those concerned with test fairness was very great indeed, although in terms of specific actions little of any consequence occurred for about one year. Evidently this constituted a period of absorbing the ultimate meaning of the developments in the theory of test fairness, and considering the practical consequences of those developments.

Things became even more confused when Nancy Cole of the American College Testing Program presented a paper offering yet another model of fairness (1972). To Cole, a better definition of fairness would be stated very simply in the following form: Applicants from different ethnic groups who would be successful if they were selected should have the same probability of being selected. If they do, then the selection process is fair; if they do not, for example in the case where a potentially successful black candidate has a one-third chance of being selected, while a potentially successful white candidate has a one-half chance of being selected, then the procedure is unfair.

Cole's model used the same assumptions and circumstances Thorndike had used, and both are concerned with relative proportions of the two applicant groups.

They differ in that Thorndike concentrates on that proportion of each group which is selected by the test, advocating that it should equal the proportion of the group who would succeed on the job. Cole, on the other hand, looks first at that part of the group which succeeds on the job, then advocates that the probability of selection for that group be the same for both minority and majority groups. Both the Thorndike and Cole models find the traditional model to be specifically unfair to minority groups, because a smaller percentage of minorities would typically be selected by that procedure than either of theirs. Thorndike's model deviates moderately, and Cole's more so, from the traditional one.

All three models sound intuitively reasonable and worthy of implementation. The trouble is, all three cannot be followed simultaneously, and in the absence of a perfectly valid test what is fair for one model is necessarily unfair according to each of the other models. It becomes clear that as long as the two selection groups in question differ on the criterion measure, there can be no single objective standard for test fairness. The traditional model is indisputably lacking in desirable characteristics, but so are the two alternatives that have been suggested, to the extent that they conflict with the traditional model and each other.

Darlington's contribution to the field, published as the second article in the journal containing Thorndike's, is the most definitive of the approaches yet made, in that his presentation incorporated all three of the competing models. Better yet, he suggests a specific means of dealing with the real-life situations.

Since there can be no single objective definition of test fairness under these circumstances, reasoned Darlington, then the only solution is to

acknowledge this fact and openly decide upon some set of values which can then be invoked by means of the selection of the appropriate numerical quantities. Specifically, Darlington offers the "corrected criterion" model as the basic format, in which the value system of the selector is used to determine the amount of correction to apply to the criterion scores for the lower-scoring group of applicants. The correction can range from very large to very small or zero, but this determination must necessarily be made on subjective, non-psychometric grounds. In effect, the Darlington model makes explicit what is only implicit in the Thorndike and Cole models, which proceeded by adopting particular new definitions of the meaning of fairness. The Darlington method of correction permits an infinite range of definitions of fairness which are explicit and open to examination and possible adjustment in response to changing conditions.

Darlington has labeled his model the "corrected criterion" because one way of implementing it is to decide on a specific increment to add to the criterion scores of the minority group. One practical means of arriving at the same effect is for the psychometrician to simply generate his estimates of the probability of success for each candidate in either group. The job of those doing the selecting then becomes one of deciding how much more risk of failure one is willing to take, if any, in order to include among those who are selected a number of the lower-scoring group who will be successful. A specific example might be that corresponding to the case described earlier, in which those members of the majority group are selected who have probabilities of one-half of succeeding on the job, while those of the minority group are selected who have probabilities of one-third. Is this an accurate reflection of the values which the selecting institution wishes to invoke? If not, how should those proportions, those risk factors, be altered to do so? The traditional

method would invoke no differential consideration at all, selecting the individual candidates with the highest probability of success regardless of group membership. The other endpoint of the range would be to consider group membership as the only consideration, selecting only the highest scoring individuals from the lower-scoring group. In such a case, the "corrected criterion" would amount to a criterion based totally on group membership. Any point between these extremes is possible depending upon the consciously chosen values of those doing the selection.

The Darlington model, then, seems to be the one most capable of accurately encompassing the intricacies of the problem, even though it necessarily offers no firmly anchored definition of fairness to which those doing the selecting can have recourse. In one sense, it amounts to an avoidance of the issue for the psychometrician, in that it removes him from the focus of attention and turns the problem over to others for solution. In another sense it represents a considerable advance in our understanding of the nature of the problem, and in addition, points up the necessity for an open and conscious examination of the value systems which are being invoked in any given selection setting. This is an important contribution in itself.

As radical as the Darlington method sounds in this context, the "corrected criterion" approach has already been used informally in many actual situations. The familiar veteran's preference in civil service selection, usually consisting of rewarding bonus points to those who served in the armed forces, is just that sort of system. Further, colleges have often "corrected the criterion" in attempting to control the available talent for the student orchestra, or to select a winning athletic team, and...

of course, the sons and daughters of heavy contributors to the endowment of the college are likely to be given more careful consideration when they apply. And although the suggestion that special consideration be given to ethnic minorities is met in some circles with a great deal of resistance, in other circles it would be agreed that, given one white and one black candidate with precisely the same qualifications, it is a greater error to turn away the black student. If that is agreed, then it is an indication that some increment of correction to the criterion, however small, is acceptable. The question then becomes one of determining the size of the increment, rather than any setting of an ominous precedent.

Ultimately, one realizes that the problem amounts to a rather special case of a problem that has plagued psychometrics eternally, that known as the "criterion problem." No one can be found who will seriously defend the freshman year grade-point average as an important gauge of anything very important in life's list of desirable values, and in fact a case is often made for its perversity. Yet, primarily because it is so easily obtained, it is the most frequent criterion variable in use for validating college selection procedures. Other things are universally agreed to be more important, which is another way of saying that the criterion is in need of being corrected, in much the same way that Darlington's model has helped elucidate. A similar correction is called for in the employment setting, where it is agreed that supervisory ratings, or even most on-the-job evaluations, leave something to be desired as criteria of success, hence are in need of some correction toward a more desirable judgment. But other, more long-range criteria (income? contributions to humanity?) are enormously difficult to measure and subject to at least as much disagreement about the need for "correction."

Added to this increased appreciation of the ethereal qualities of what we are attempting to predict is the very practical realization that in a great number of selection settings, the total size of the operation is simply too small to apply even these imprecise and ultimately subjective methods to the process. However, rather than conclude from these travails that we have engaged in a useless exercise, let it be pointed out that the ultimate solutions are the same, both in large samples and small, and rest with the values of the selector. In fact, it is comforting to realize that the current legal inducements to the hiring of minorities are pointed in essentially the same direction that our very elaborate psychometric acrobatics, just described, would have us go. We have made advances, specifically in that we no longer adhere to a traditional model which was thought to be the ultimate definition of fairness, but in fact was not so fixed and unquestionable after all. The battle for fairness is by no means over as a result, but at least the lines have now been more clearly drawn.

Given this present understanding, and by way of review, what does the selector, either the admissions officer or the employment director, do in the real-life decisions that he must make? From a psychometric point of view, the procedure is quite straightforward and not greatly different from existing procedures. The probability of success of each applicant is obtained from the psychometrician, based on the performance of similar-scoring candidates of previous years, and on ethnic identity. The selector then chooses his group of "admits" basing his decisions on this information and the value he places upon the selection of a particular number of minority applicants. In any given circumstance, he may be required to utilize two different estimated probabilities of success in order to invoke these values, frequently taking a higher risk of failure with the minority group applicants.

There are additional points for consideration, to be sure, which may make the selector's task more difficult. If the number of minority applicants changes dramatically from one selection to the next, perhaps in response to increased recruitment activities, for example, then this is likely to alter the success rates and confuse the selection process, especially if it has been based upon some assumption about the success-rate in the applicant population (as is the case with the Thorndike or Darlington models). Further, the impact of the changing student body, or employee group, upon the manner in which the criterion behavior itself is performed (is a school changed by its students?), is likely to require constant updating of the success-rates as well as the contents of the selection test battery itself.

If these difficulties are overcome, then there may still be problems encountered by the selector as a result of the currently confused legal status on this same problem. Although federal "affirmative action" programs would appear to conform rather completely to the psychometric conclusions described above, there is a characteristic of these procedures which make them similar to the implementation of a "quota system," a concept which possesses quite negative associations for many people, and has in fact been specifically denounced by President Nixon. The apparent contradictions between "affirmative action" and "quota system" have yet to be resolved. Meanwhile, the Equal Employment Opportunity Commission has adopted a policy for implementation which conforms most closely with the traditional model and which, in particular, requires that separate validation procedures be employed for minority groups.

This ruling was intended to increase the incidence of minority employment. It was based on the early belief that the use of a single prediction equation for both minority and majority groups would be unfair to minorities. It is therefore surprising to encounter considerable empirical evidence suggesting that in the typical selection setting it is more often the case that the performance of minority group members on the criterion is over-predicted by the use of a single prediction system. The ultimate effect of that EEOC guideline would be to eliminate this overprediction, thus ironically eliminating the small inadvertent bonus to minorities that had been inherent in the prevailing system. Some evidence (Linn, 1973) indicates that the degree of this overprediction is nearly but not quite sufficient to satisfy the Thorndike definition of fairness. At any rate, the elimination of that overprediction can hardly be considered a solution. Rather, its empirical effect of causing the selection of fewer minority group members should be acknowledged and dealt with.

Thus, the practical consequences of the developments described here are that the values held by those doing the selecting must ultimately be invoked in the selection process, as indeed they frequently have been in the past, but they now need to be described in objective terms. These objective terms, taking the form of particular levels of acceptable success likelihood for any subgroup, are the material from which a fair selection process must be built.

Also on the practical level, the modest size of a great many selection operations, especially in industry, is simply insufficient to permit the use of the statistical tools that provide such objectively precise forecasts of success. However, since it is now seen that subjective values must

guide the decision anyway, perhaps there is less reason to attempt to approximate the large-sample models often invoked as the only means to fair selection. Abandoning the attempts to imitate such an inappropriate model might well result in fewer real selection errors.

Still another consideration in the attempts to effect the optimal selection system must be that of the consequences of those decisions, both those of rejection and those of acceptance, and this should interact with the invoked value system, if not made an integral part of it. The consequences of an error will vary from setting to setting. For example, between selection for employment in a high-cost or high-risk position, where a wrong selection decision might have very serious results, in contrast to a selection decision for an educational opportunity in which the consequences of a wrong decision might not be so great.

Meanwhile, two problems remain which may interfere with the implementation of the methods suggested here. First, once selection is made differentially on the basis of ethnic identity, then other identifiable subgroups could reasonably request special treatment as well. The population is capable of infinite subdivision and cross-classifications that could be invoked, with sex and socioeconomic status being the most likely next divisions. It can be anticipated that disputes and difficult decisions will surround these topics in the future.

The second problem with the procedure suggested here is that at least for now, the Supreme Court is on record as being opposed to it. In Griggs vs. Duke Power, as Campbell (1973) has pointed out, the decision states that "Congress has made [job] qualifications the controlling factor so that race, religion, nationality, and sex become irrelevant." Making

race irrelevant, of course, would specifically prohibit the suggested procedure. More recently, however, the state Supreme Court of Washington has ruled, in DeFunis vs. Odegaard, that racial distinctions can be made for compensatory purposes, at least in educational selection practices (Jones, 1973).

Clearly all the problems are not yet solved, but it is now obvious that further attempts at establishing a final and fixed standard, derived from the psychometric characteristics of the problem, are not appropriate.

References

Campbell, J. T. Sources of bias in the prediction of job performance, final report. Princeton, N. J.: Educational Testing Service, 1973, in press.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cole, N. S. Bias in selection. ACT Research Report No. 51. Iowa City, Iowa: American College Testing Program, 1972.

Darlington, R. D. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Jones, R. F. (Ed.). Washington Reports, 82 Wn 2d 1-73, 1001, No. 1, March 30, 1973, DeFunis vs. Odegaard, p. 11-68.

Linn, R. L. Task Force B - Background Paper on Test Bias, unpublished ms. dated 1-16-73, Educational Testing Service, Princeton, New Jersey.

Thorndike, R. L. Concepts of culture fairness. Journal of Educational Measurement, 1971, 8, 63-70.